Assessment and Feedback: Examining the Relationship Between Self-assessment and Blind Peer- and Teacher-assessment in TOEFL Writing

Meghan Odsliv Bratkovich[1]

ABSTRACT

This study investigated the nature of self-assessment and blind peer- and teacher-assessment in L2 writing. The type of feedback students gave to themselves and peers, the type of feedback used in the revision process, and the source of the feedback used were all analyzed. Additionally, student perceptions of self- and peer-assessment, feedback, and their relationships to perceived writing improvement were also studied. Findings revealed that students in this study did not use teacher feedback significantly more than feedback from themselves or their peers, but they did give different types of feedback than the teacher and favored using feedback related to language use in the revision process. Students perceived their writing abilities to have increased due to self- and peer-assessment but responded more positively to peer-assessment than self-assessment. Surprisingly, students also perceived their abilities to have increased in rubric areas in which the feedback they received was not used and not regarded as useful, and the highest perceived gains in writing ability were in areas which accounted for the lowest amounts of feedback given.

INTRODUCTION

Second language (L2) classroom teachers have long been interested in improving student writing. One of the primary ways in which teachers help their students improve is through assessing and giving feedback on their written work. The culture of assessment around the world, however, has been slowly shifting from a more summative approach to a more formative one. Since Bloom, Hastings, and Madaus (1971) first contrasted *evaluation* and *assessment*, separating the judgment associated with summative evaluation from the engagement of teaching and learning associated with formative assessment, educators and researchers have been exploring formative assessment methods. Though summative tests certainly still have their place in education around the world, it is formative assessment that "offers great promise as the next best hope for stimulating gains in student achievement" (Cizek, 2010, p. 3). Two of the many formative methods that have garnered attention are self-assessment and peer-assessment, which are the focus of this study.

[1] Meghan Odsliv Bratkovich received her EdM degree in Applied Linguistics and is currently completing her PhD in Teacher Education and Teacher Development at Montclair State University in New Jersey. She can be reached at meghanodsliv@gmail.com.

LITERATURE REVIEW

Self-assessment is defined by Luoma (2013) as "the language learner's evaluation of his or her own language skills, usually in connection with a language course or as part of other forms of language assessment" (p. 1). Peer-assessment, on the other hand, is the "complement to self-assessment" (Black, Harrison, Lee, Marshall, & Wiliam, 2004, p. 14) and is defined by Topping (2009) as "an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal status learners" (p. 20).

Although self-assessment has attracted attention in fields of self-regulation and formative assessment, it could easily be argued that it has always been an integral component of good teaching and learning. Self-assessment in language education gained momentum through initiatives by the Council of Europe (1981, 1988) and early learning-oriented frameworks such as Chamot and O'Malley's (1994) cognitive academic language learning approach (CALLA). Both peer- and self-assessment were also essential components of the process approach to writing (Elbow, 1973), which emerged in the 1970s in first language (L1) writing and rose in popularity in the 1980s and 1990s to also include second language writing instruction. More recently, peer- and self-assessment have played significant roles in classrooms adopting collaborative learning approaches, as peer-assessment allows students to help each other in the learning process.

Benefits of peer-assessment and self-assessment

The rise of communicative language teaching and more formative approaches have changed instructional goals to become more communicative and standard-based than in previous years, which has, at least theoretically, led to more transparent and comprehensible goals. These goals are easier to conceptualize, and a likely effect is better awareness of what is to be studied and why (Oscarson, 2014, p. 713). This awareness is essential given that self-assessment can only occur when the students "have a sufficiently clear picture of the targets that their learning is meant to attain" (Black & Wiliam, 1998, p. 5).

From a formative assessment standpoint, peer- and self-assessment have become increasingly popular as learning tools, as they encourage the development of metacognitive skills such as identifying strengths and weaknesses and planning future learning. Peer-assessment requires students to reflect, intelligently question, and make judgments, which can then promote self-assessment and self-awareness (Topping, 2005; Topping & Ehly, 1998). MacArthur (2007) similarly claims that the revision process after peer-assessment may not only improve students' current pieces of writing but also improve their general writing ability and their ability to self-assess their own works.

Hansen Edwards (2014) notes the numerous cognitive and metacognitive benefits of peer- and self-assessment, citing the increased time spent thinking, reviewing, and summarizing—all of which lead to the development of autonomy and greater understanding of high quality work, the nature of writing, and the assessment process. Peer-assessment has been associated with gains for both assessors as well as assesses in L1 contexts (Topping, 2005; Topping & Ehly, 1998), gains that Topping (2009) attributed to increased levels of practice, time on task, sense of accountability, and the possible identification of knowledge gaps. Topping (2009) also noted that "cognitive and metacognitive benefits can accrue before, during, or after the peer-assessment. That is, sleeper effects are possible" (p. 23). Although Hansen Edwards (2014) named potential drawbacks of peer-assessment, specifically with respect to time,

perception, and quality of feedback, she listed no potential cognitive or metacognitive disadvantages to peer-assessment.

Peer- and self-assessment within a L2 context have been used in much the same way as in other disciplines, with peers assessing a variety of oral and written work. Although research in second language classrooms is more limited, studies echo similar cognitive benefits as in content classrooms, as peer-assessment helps students "take charge of their own learning, build critical thinking skills, and consolidate their knowledge of writing" (Liu & Hansen, 2002, p. 1). Liu and Hansen (2002) state that peer-assessment enables L2 students to understand their own drafts better and provides guidance for revising content. Liu (1997) found many advantages across an array of levels that L2 students perceived from peer-assessment. On the textual level, students felt better able to recognize their own errors and identify weaknesses in their drafts. On the cognitive level, students reported better idea organization and improved critical thinking, and on the communicative level, students said peer-assessment provided good opportunities to both express opinions and listen to those of others.

In a study experimenting with various combinations of self-, peer-, and teacher-assessment in English as a Foreign Language (EFL) writing classes, Birjandi and Hadidi Tamjid (2012) found that the group using self-assessment paired with teacher-assessment performed significantly better on the post-test than the group using only teacher-assessment. Similarly, the group using peer-assessment paired with teacher-assessment performed significantly better than the group using only teacher-assessment. These findings provide evidence that self-assessment in addition to teacher-assessment, and peer-assessment in addition to teacher-assessment, yield greater improvement on writing performance than teacher-assessment alone. The combination of peer- and self-assessment, however, did not yield significantly different scores as compared to teacher-assessment alone, indicating that teacher-assessment is perhaps a primary factor in writing improvement in this context. Birjandi and Hadidi Tamjid conclude that both peer- and self-assessment are advantageous in an EFL writing classroom and attribute their findings to the shared responsibility for the management of learning, self-directed learning, and learner-centered teaching.

Though Birjandi and Hadidi Tamjid (2012) provided rich and sound quantitative data, they did not address the combined effect of all three types, peer-, self-, and teacher-assessment. Given that no significant relationship was found between the combination of teacher-/peer-assessment and teacher-/self-assessment, studying all three together could provide some insight into whether peer-assessment, self-assessment, or both, when combined with teacher-assessment, might yield the greatest gains.

While feedback generated from peer- and self-assessment, as well as teacher-assessment, has many potential benefits, even if given in copious amounts, this feedback does not automatically cause an improvement in the quality of writing performance (Hansen Edwards, 2014). While London and Tornow (1998) note that feedback from multiple perspectives promotes self-awareness, it is noticing the gap between self and others' perception which is a probable factor leading to further learning and improvement in writing quality.

Both peer- and self-assessment fit within a formative framework of assessment that seeks to assess "the acquisition of higher-order thinking processes and competencies instead of factual knowledge and low-level cognitive skills" (Lindblom-Ylänne, Pihlajamäki, & Kotkas, 2006, p. 51). Despite the popularity, advantages, and theoretical support of peer- and self-assessment, due to the high variation in practices, there is not yet consensus on what constitutes effective peer- or

self-assessment, which measures lead to increased student learning, and no overarching theory or model of the process has seemed to emerge in either peer- or self-assessment.

The nature of feedback in peer-assessment and self-assessment

Subject-matter experts and novices, or teachers and students in the contexts of these studies, may generate very different feedback due to the different domain-specific knowledge, schema, and problem-solving abilities of each. Since students are novices in their disciplines, they do not yet have the extensive knowledge and skills of a seasoned expert, which could limit their ability to provide helpful feedback. Though students may perceive themselves and their peers as giving lesser-quality feedback than teachers, Topping (1998, 2003) suggests that there is little difference between the quality of teacher as opposed to peer feedback, and the teachers in Weaver's (1995) study actually found feedback from peer responses to be more effective than their own. Additionally, peer-assessment generally yields a greater amount of feedback than teacher-assessment (Hyland & Hyland, 2006), giving students more information about their performance. While Jacobs and Zhang (1989) found that teachers provide more accurate feedback in the area of grammatical accuracy, peers provide feedback on informational and rhetorical accuracy with similar quality to that of teachers. Cho and MacArthur (2010), referencing L1 students, also postulate that feedback from peers may actually be better understood due to shared knowledge and difficulties. An interesting finding in Lindblom-Ylänne et al.'s (2006) study, which was only briefly mentioned, was that students reportedly felt that it was easier to assess technical aspects of writing than aspects of content. Though the researchers did not offer an explanation, this could be related to expert versus novice knowledge and cognition within a particular subject domain.

A problem that has been observed in both L1 and L2 peer-assessment is that students tend to focus their feedback on surface-level revisions such as spelling, vocabulary, and grammar rather than on deeper-level revisions such as organization and idea development. Beason (1993), for example, found that L1 writers in content-area courses primarily addressed surface-level concerns, and Yagelski (1995), noticing a disconnection in the relationship between classroom context and the nature of revisions, drew similar conclusions. Leki (1990) and Villamil and De Guerreo (1998) had comparable findings with L2 students, commenting that instead of actively engaging with texts and responding to the meaning conveyed, peer assessors "are likely to respond to surface concerns of grammar, mechanics, spelling, and vocabulary, taking refuge in the security of details of presentation rather than grappling with more difficult questions of meaning" (Villamil & De Guerreo, 1998, p. 9). Interestingly, Villamil and De Guerreo (1998) and Connor and Asenavage (1994) found revisions made from teacher feedback similarly led to only surface-level revisions. Though there are likely ample reasons for this pattern of student behavior, Villamil and De Guerreo (1998) postulated that, due to learning gaps in language structure, students in their study felt the need to first address aspects of form and fix linguistic errors, or perhaps, due to previous language teaching focused on attention to grammatical form, simply fell back on old habits of learning. Liu and Hansen (2002) further add that a lack of confidence in pointing out content-based flaws may also contribute to the pattern.

Second language writing teachers generally agree that the most helpful comments in peer-assessment are those that address global issues such as content and organization (Liu & Hansen, 2002), yet students tend to follow their teacher's lead and comment on areas in which

their teacher usually comments (Liu, 2013), meaning that if the teacher comments mostly on grammatical accuracy, then so will the students.

In a study exploring the relationships between self-, peer-, and teacher-assessment of student writing in a content-area course, Lindblom-Ylänne et al., (2006) incorporated not only qualitative feedback but also rating procedures against a scoring matrix. In this case study, student essays were self-assessed, then blindly rated by both an instructor and a peer. Findings showed that overall self-, peer-, and teacher-ratings were quite similar, but that while peers were more critical on some aspects of the rubric, teachers were more critical on others—the largest disparity being in the area of independent thought. Lu and Bol (2007), examining anonymous computer-mediated review, found that students who participated in anonymous peer-assessment not only gave more critical feedback to peers, but also showed more writing improvement than students whose identity during peer-assessment was known.

Saito and Fujita (2004), in one of the few studies to focus on student rating rather than feedback, examined ratings of peer-assessment compared to those of self-assessment and teacher-assessment. They found that while ratings from teachers and peers were strongly correlated, ratings from self-assessment had no correlation with either peer- or teacher-assessment. Saito and Fujita (2004) concluded that "self-rating is idiosyncratic and strongly contingent on a subjective view of self-product" (p. 47).

Matsuno (2009) found that Japanese student raters self-rated more harshly and peer-rated more leniently than expected. She notes that this tendency was independent of the ability level of the writer, but acknowledges that Japanese cultural factors such as humility and group harmony likely contributed to her findings. Similarly, though students in Lindblom-Ylänne et al.'s (2006) study found it difficult to be critical of peers during peer-assessment, self-assessment proved to be more difficult because of the perceived impossibility of being objective when self-assessing. Their tendency was to be overly critical, a finding that is consistent with Matsuno's (2009) results but does not support Sullivan and Hall's (1997) and Falchikov and Boud's (1989) findings claiming that students tend to self-rate themselves higher. As MacLeod (1999) found, interpersonal relationships can certainly alter the content of peer-given feedback, as reviewers of a variety of ages, nationalities, and content areas tend to rate higher and provide less critical feedback to their peers, likely in an effort to preserve relationships. As Zhao (1998) found, however, anonymous peer-assessment conditions led to more objective ratings.

How students use self-, peer-, and teacher-given feedback during the revision process could be different in L1 and L2 contexts. While Black, Harrison, Lee, Marshall, and Wiliam (2003) posit that L1 students more readily accept feedback or criticism from a peer than a teacher, this has not been found in L2 research. Conversely, Liu and Hansen (2002) assert that students more carefully attend to teacher-given feedback than peer-given feedback, and Zhang (1995) found that L2 students appeared to use teacher-given feedback more so than feedback from peers. Similarly, Cheong (1994) found that while high school L2 students incorporated feedback from self, peer, and teacher sources in their revisions, they mostly used feedback from teachers in the revision process. In addition to using feedback from teachers more often than feedback from other sources, L2 students report the most positive attitudes toward and preferences for teacher feedback as well. Zhang (1995) and Nelson and Carson (1998), in studies comparing student preferences for feedback, found that students most prefer teacher feedback, followed by feedback from peers, and finally self-given feedback is least preferred. Since the sources of the feedback in these studies were apparent to the students, however, it is not clear if feedback was used or preferred due to its perceived quality or the status of its source.

While many studies have described the nature of feedback and how it is used in student writing, fewer studies have examined how students perceive peer- and self-assessment; however, those that have cite generally favorable attitudes. Foley's (2013) students viewed peer-assessment as an overall positive experience, and Ballantyne, Hughes, and Mylonas (2002) found that students reported that peer-assessment benefited their learning process. While Foley's (2013) students reacted approvingly to peer-assessment, their praise was somewhat tempered due to the amount of classroom time lost and skepticism regarding the quality of feedback given by peer assessors. Saito and Fujita (2004) also found that attitudes toward peer-assessment were not influenced by peer ratings as students receiving low peer scores viewed peer-assessment as favorably as those receiving high peer scores. Surprisingly, students in Foley's (2013) study also perceived that the primary benefits of peer-assessment lie with the assessor rather than the assessee.

Conclusions drawn

The combination of internal self-assessment with anonymous external assessment from a novice peer and expert teacher has the potential to provide a unique and triangulated view of L2 student writing. While multiple studies have examined peer- and self-assessment in a variety of contexts, fewer studies have researched the combination of self-, peer-, and teacher-assessment, and none of these studies have addressed second language writing contexts. Additionally, while a few studies have utilized anonymous or blind peer-assessment, no studies found utilized blind teacher-assessment in a L2 context. Since students are generally believed to favor teacher feedback over that of a peer, no studies have attempted to answer if this preference is due to feedback quality or the status of the assessor. Furthermore, studies addressing student attitudes toward feedback from anonymous sources could not be found.

PURPOSE OF THE STUDY

The purpose of this study is to further understand how students view and respond to self-assessment and blind peer- and teacher-assessment in a writing context. A few key terms and concepts are referred to throughout the study. Firstly, within the context of this study, feedback refers to the narrative information from self, peer, and teacher sources that gives comments on, or suggestions for, how a piece of writing could be improved. Secondly, blind peer-assessment in this study means that students were not aware of the identity of either the person who assessed their essay or the person whose essay they assessed. Additionally, the feedback from peer and teacher sources was also given blindly, meaning that the identity of the source of the feedback that was given to the original writers was not identified as being either from the teacher or a peer. When students completed the revision process, feedback use was examined. Within this study, feedback use refers to evidence that the feedback given had been incorporated into the writing, regardless of whether or not it actually improved the piece of writing.

This paper contains the method, results, and conclusions for a small study examining self-assessment, and blind peer- and teacher-assessment, in a L2 writing context. First, the research questions are presented; then a detailed method section describing the research design, participants, materials, data collection procedures, and analysis follows. The results of the analyses are then discussed, followed by conclusions drawn.

RESEARCH QUESTIONS

1. What is the nature of feedback source and feedback use among self-given feedback and blind peer- and teacher-given feedback?
2. What is the nature of student perception of improvement as it relates to rubric areas, self-given feedback, and blind peer- and teacher-given feedback?

METHOD

This section of the paper will address the method of the study. After schematizing the research design, the participants, instruments, and materials will be described. Data collection and their subsequent analysis will conclude this section.

Research Design and Study Variables

This study will use a single group time series design. The primary dependent variable in this study is writing ability, operationalized by scores on a TOEFL® writing prompt. Additionally, student-generated feedback and feedback response are also treated as dependent variables, and are measured according to the frequency and type of feedback generated or used. The design used for this study is schematized below:

G_1—X_1—O_1—X_2—O_2—X_3—O_3—X_4—O_4—X_5—O_5—X_6—O_6—X_7—O_7—X_8—O_8—X_9—O_9—O_{10}

Participants

Participants in this study were seven students enrolled in the Community English Program (CEP) at Teachers College, Columbia University. The CEP is an English language program administered by the Applied Linguistics and Teaching English to Speakers of Other Languages (TESOL) department. The program provides English as a Second Language (ESL) instruction to adult learners from a variety of ethnic and cultural backgrounds who are living in the greater New York City area. ESL classes are taught at Beginner (B), Intermediate (I), and Advanced (A) proficiency levels, with each level typically comprising four sub-levels, i.e., B1, I3, A2, etc. Based on their scores on the CEP Placement Exam, students are placed into the level that best matches their overall proficiency. Though most classes within the CEP are for general ESL using an integration of skills, these student participants were enrolled in a specialized TOEFL preparation course and were purposefully sampled based on their participation in this course. To enroll in the TOEFL prep course, students must have tested at the I4 level or higher, roughly comparable to at least a high intermediate level.

The student participants, henceforth referred to as the students, were six women and one man, and represented four native language backgrounds: Turkish, Russian, Japanese, and German. Most were between the ages of 25 and 35 and had been studying English for at least seven years. With the exception of one student who was taking the TOEFL to be admitted to an undergraduate program, all students had completed university in their home countries.

The researcher, henceforth referred to as the teacher, was a native English speaker, an experienced ESL teacher, and graduate student in the Applied Linguistics program at Teachers

College, Columbia University. The classroom teacher for the course, a second rater, and a second coder were non-native though highly proficient English speakers who were also experienced ESL instructors and Applied Linguistics graduate students at Teachers College. The teacher, classroom teacher, second rater, and second coder were all familiar with the TOEFL exam as all four had taught TOEFL preparation skills in the past, and the three non-native speakers had previously taken the TOEFL exam; none, however, had any formal training in scoring the TOEFL exam.

Instruments and Materials

The TOEFL writing prompts that were given to students each week were taken from both ETS as well as external sources. While all the independent prompts used were directly from the ETS website, the integrated prompts used for this study were from various TOEFL preparation study guides that included audio recordings. According to ETS (2011), the writing section of the TOEFL has a reliability estimate of 0.74 and a standard error of measurement of 2.76. While the writing section has the lowest reliability of the four TOEFL sections, such a reliability measure is typical for writing measures consisting of only two tasks (Breland, Bridgeman, & Fowles, 1999; ETS, 2011).

The TOEFL writing rubrics for both integrated and independent tasks were also used in this study. The holistic rubrics use a scale from 0 to 5 points and describe student writing at each interval using a variety of characteristics. For independent tasks, scoring intervals are described through use of language, organization, addressing the topic, and explanation and elaboration. For integrated tasks, scoring intervals are described through use of language, organization, presenting main points, accuracy, and integration. A few sample essays from the ETS website, along with their score explanations, were also given to the students.

The students also used self- and peer-evaluation forms created by the teacher. The self-evaluation form included the TOEFL rubric designed for the respective task type and asked students to use the TOEFL rubric to identify the strongest and weakest aspect, score the essay, and provide one or two ways in which the essay could be improved. A sample self-evaluation form can be found in Appendix A. The peer-evaluation form included all the elements of the self-evaluation form, but also included a section in which specific grammatical errors could be pointed out. A sample peer-evaluation form can be found in Appendix B.

At the end of each week, students revised their essays using a revision evaluation form which asked students to use a track changes feature or a different color font to indicate what they changed or added in their essays. The revision evaluation form also asked students to indicate their impression of the feedback they were given, simply by *liked* or *didn't like*, as well as whether or not they used each piece of feedback in their revision. This revision evaluation form can be found in Appendix C.

Lastly, at the conclusion of the study, students completed an anonymous online survey designed to elicit more qualitative data regarding their experiences. The survey was created by the teacher in response to findings from other similar studies. Some questions allowed for open-ended responses, such as describing the positive and negative aspects of the evaluation process, while others involved using a Likert-scale, for example, questions indicating to what extent students felt their writing improved in certain areas. The complete list of survey questions can be found in Appendix D.

Data Collection Procedures

The procedures followed in the current study will be examined in this section of the paper. The specifics of the treatment will be explained, as well as how data were collected, handled, and analyzed.

Treatment procedures

The primary treatment in this single group time series study was the repetition of writing, self-assessing, peer-assessing, and revising a series of essays over the course of a ten-week TOEFL preparation course. The class met once a week for three hours for a total of 30 classroom instructional hours. Though the course in question was designed to cover all four sections of the TOEFL exam, so as to prevent the confounding of variables with writing instruction, no explicit writing instruction was given by the classroom teacher. Instead, students were asked to participate in this study in lieu of writing instruction, and their participation was factored into their homework grade. Though no explicit writing instruction took place during the course, it should be noted that instruction in grammar and vocabulary was provided, as was instruction on general testing strategies and the TOEFL speaking section, which, except for the medium, shares many similarities with the writing section.

During the first class meeting, the teacher and classroom teacher introduced the study to the students, provided instruction in how to participate in the study, and explained that all writing practice and instruction for the course would be done online through this study. The students then received copies of the TOEFL writing rubrics, and the teacher conducted a brief norming session which involved detailing the elements of the TOEFL rubrics, reading a few sample essays, and rating them accordingly as a class.

The writing and assessment portion of the study began in the second week of the course. Every week of the subsequent nine weeks of the course, students received either an independent or integrated writing prompt in a document file via email from the teacher. In weeks 2, 3, 6, 7, and 10, students received an independent prompt, and in weeks 4, 5, 8, and 9, students were given an integrated prompt. Integrated prompts also included a link to an online reading passage and audio recording.

Each student responded to the prompt within the TOEFL time restrictions by typing their responses into the document file for the given prompt, completing the self-evaluation form, and emailing the document with the essay and self-evaluation form back to the teacher. After collecting all initial essays for the respective week, the teacher then redistributed the essays for peer-assessment. Each student then received an email containing a peer's essay on the same prompt, a copy of the TOEFL writing rubrics, and a peer-evaluation form. After reviewing and providing feedback on a peer's essay, each student completed the peer-evaluation form in a document file and submitted it via email back to the teacher. The teacher also completed peer-evaluation forms for each student essay.

After receiving all peer-evaluation forms, the teacher compiled the feedback for each student's self-evaluation, peer-evaluation, and teacher-evaluation into one document and sent all the feedback back to the original student. Students were then asked to revise their essays using a track changes feature common to many word processing programs and use the revision evaluation form to indicate their impression and use of the given feedback. Students then emailed their revision and evaluation form back to the teacher.

This process of writing, self-assessment, peer-assessment, and revision spanned one week and was repeated for a total of nine treatments over nine continuous weeks. At the conclusion of the study, the link to the online survey was emailed to the students for their completion.

Data collection, coding, and handling procedures

Throughout the study, steps were taken to ensure complete anonymity for the students and the integrity of the blind rating system. All emails to the group used the Bcc feature so email addresses could remain anonymous. Upon receiving each week's initial essays from students, any identifying features or names within the documents were removed both to ensure anonymity and avoid potential contamination by influencing the peer response. The files were renamed using a numeric coding system, and then converted into a portable document format (pdf) so any automatic grammar or spelling alerts common to many word processors would not influence feedback or scoring.

All reviews were done blind, meaning that it was never disclosed which other student received someone's essay, or to which other student someone's essay was given. Reviewers were systematically rotated so that over the course of the study, each student reviewed and was reviewed by each peer at least once.

When feedback was compiled before it was returned to the student for revision, formatting of corrections and feedback were standardized so as not to indicate which feedback was associated with which source. Although it was assumed that students would be able to identify feedback from their own self-evaluation form, the feedback from the teacher and the peer was identical in format and presented in varying order so that students could not notice patterns characteristic of either peer- or teacher-given feedback (e.g., peer feedback is always last, or teacher feedback always points out errors using red font). Not providing information as to which feedback came from the teacher and which came from the peer was done to greater ensure that revisions based on feedback were made due to the content and quality of the feedback rather than the status of the reviewer.

Grammar corrections pointed out in the fourth section of the peer-evaluation form by peer and teacher evaluators remained in that section and were passed on to the student, but the suggestions for improvement were coded and categorized. At the conclusion of each week, the teacher examined the suggestions for improvement in the feedback evaluation given to the students against their first drafts and their revisions, looking for evidence of incorporated feedback. The suggestions for improvement were then coded and categorized by their use or disuse, student impression (liked or didn't like), and source of the feedback (self, peer, or teacher). It is important to note that students often over-reported their use of feedback, so all feedback evaluation was scrutinized for any evidence of incorporation. *Use* was operationalized as any evidence that feedback had been incorporated into the writing, regardless of whether or not it actually improved the piece of writing. If, for example, feedback was given that more explanation in the second paragraph was needed, and an additional descriptive sentence was added in the second paragraph, this was counted as use even if it did not improve the quality of the paragraph. Similarly, if feedback was given that more connectors would improve the organization, but there was no evidence of the addition of any transition words or connecting language, it was categorized as unused.

Though data attempting to ascertain whether or not students liked the feedback given to them were collected, in only a select few instances did students report that they disliked the

feedback, usually further noting that they did not understand the feedback that they did not like. Additionally, some feedback was simply ignored by students and never labeled liked or disliked. Because of the extremely low reporting rates of feedback that was not liked, and some ignored feedback, the like/dislike differentiation was eliminated and all suggestions for improvement were simply coded and categorized as either used or unused, and from a peer, self, or teacher source.

Additionally, all suggestions for improvement in the feedback provided were further coded according to the aspect of the rubric to which it pertained. Suggestions for improvement were thereby coded according to use of language, organization, addressing the topic, and explanation and elaboration for independent tasks, and according to use of language, organization, integration, presenting main points, and accuracy for integrated tasks. One additional category, *Unspecific feedback*, was added. Unspecific feedback was most often a strategy such as *look at more examples*, *study grammar more*, or *use time more effectively* or simply feedback that was not specific enough to lead to any likely changes. Feedback such as *improve grammar*, for example, was categorized as unspecific unless it was accompanied by a specific grammatical aspect that the writer could address. Several students suggested using more sophisticated vocabulary, and although this is not an incredibly specific suggestion for improvement, some students did change some vocabulary words in their revisions, so suggestions for more sophisticated vocabulary were categorized with use of language. Similarly, the limited amount of feedback addressing punctuation, such as commas and quotation marks, and other issues of mechanics was also coded as use of language.

Due to slight variances in the language of the integrated and independent TOEFL writing rubrics, some similar feedback was coded differently based on the task type. The addition of detail, for example, was coded as explanation and elaboration in independent tasks, while in integrated tasks it was categorized as accuracy. The reason for this is that the scoring rubric for integrated tasks does not contain language explicitly evaluating explanation and elaboration. The concept of fully developing one's idea in an integrated prompt, therefore, falls under the accurate presentation of information. Similarly, issues relating to staying on topic were categorized as addressing the topic in independent tasks, but as presenting main points in integrated tasks.

Scores from 0 to 5 given by self, peer, and teacher sources were also recorded each week. While students were encouraged to assign only one number, some students could not decide and gave a range, double scores, or a half score. While this is not practiced among trained TOEFL raters, in this study ranges, such as *2-3*, and double scores, such as *3 or 4*, were averaged to produce a half score between the two for analysis purposes, and half scores were maintained. All suggestions for improvement, their respective coding, and scores from self, peer, and teacher sources were recorded in a spreadsheet each week.

Data from the survey taken at the end of the study were also categorized and coded. Open-ended essay responses were organized by question and coded according to their respective aspect of the writing rubric. Likert-scale responses were converted to numerical data by assigning a point value from 1 to 5 (for questions with 5 variations in the response) or from 1 to 4 (for questions with 4 variations in the response) in preparation for analysis.

Data Analysis

All data were analyzed in IBM SPSS Statistics 21. To begin, assumptions involving normality were tested and descriptive statistics were analyzed according to source (self, teacher,

or peer). Mean scores and standard deviations were calculated for self, peer, and teacher sources, and preliminary analyses revealed low skewness and kurtosis values indicating normal distribution and allowing for further analysis.

To provide evidence for consistency in measurement of the teacher's scores, both inter-rater and inter-coder reliability between the teacher and an independent rater/coder were calculated for the scores and codes assigned in weeks 2, 6, and 10, representing a third of the total sample. These weeks were selected because they represented the beginning, mid-point, and conclusion of the study. Because the scoring was on an ordinal scale, the Spearman rank-order correlation procedure was used to estimate the correlation coefficients for inter-rater reliability. Correlation coefficients were also calculated for self, peer, and teacher scores for all weeks of the study. Inter-coder reliability with the categorical feedback data was calculated using Cohen's Kappa.

To determine if there were any significant gains for the group as a whole, paired sample t-tests on all three sets of scores (self, peer, and teacher) were used to analyze the scores at the beginning and end of the study. A repeated-measures ANOVA was also analyzed to provide further insight into the improvement of writing scores. Assumptions for both normality and sphericity were met through insignificant Mauchly's Test of Sphericity, and the low skewness and kurtosis values previously determined. Data relating to the type of suggestions for improvement in feedback and its use or disuse were then analyzed using chi-square tests. Chi-square assumptions regarding expected counts and frequencies were met. Lastly, linear regression analysis was used to determine if the use rates of different types of feedback significantly increased or decreased throughout the study. Assumptions of linearity and homogeneity of error, demonstrated by an insignificant lack of fit and Breusch-Pagan tests, were met.

To analyze survey results, descriptive statistics and measures of central tendency on all Likert-scale questions were performed. A repeated-measures ANOVA was analyzed to determine if students perceived their abilities to have improved to different extents across the writing rubric. Paired samples t-tests were then run comparing student attitudes toward peer-assessment and self-assessment. Assumptions for both normality and sphericity were met through insignificant Mauchly's Test of Sphericity and low skewness and kurtosis values. Open-ended questions were examined to provide qualitative descriptions of statistical results.

RESULTS AND DISCUSSION

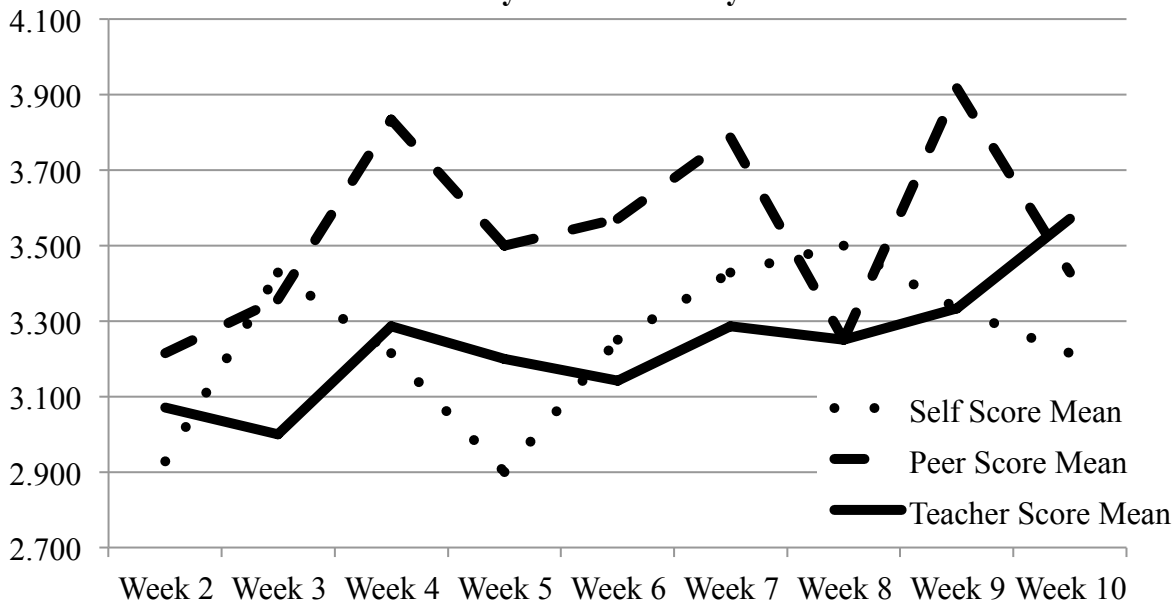
Descriptive statistics calculated included the mean scores and standard deviations of the scores assigned to the essays of seven students (N=7) over nine total weeks (weeks 2 through 10 of the course). Scores were divided into assignment of score by source: either self, peer, or teacher. Table 1, below, summarizes the mean scores and standard deviations.

TABLE 1
Descriptive Statistics by Source

	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
Self Score Mean	2.929	3.429	3.214	2.9	3.25	3.429	3.5	3.333	3.214
SD	1.17	0.787	0.699	0.742	0.418	0.607	0.577	0.516	0.567
Peer Score Mean	3.214	3.357	3.833	3.5	3.571	3.786	3.25	3.917	3.429
SD	0.994	0.748	0.983	0.5	0.535	0.809	0.5	1.021	0.787
Teacher Score Mean	3.071	3	3.286	3.2	3.143	3.286	3.25	3.333	3.571
SD	0.732	0.816	0.756	0.447	0.627	0.951	0.5	1.033	0.607

The data seem to indicate that self-assigned scores were somewhat inconsistent and that peer-assigned scores were generally higher than either self- or teacher-assigned scores, findings consistent with Lindblom-Ylänne et al. (2006), Matsuno (2009), and Saito and Fujita (2004). The mean scores for each source were charted below in Figure 1 to better visualize the changes.

Figure 1
Weekly Mean Scores by Souce



The results of the paired samples t-tests comparing scores at the beginning and end of the study within self-assigned, peer-assigned, and teacher-assigned scores indicated that scores at the beginning and conclusion of the study were not significantly different for all three sources. Similarly, repeated-measures ANOVA results were insignificant for self-assigned, peer-assigned,

and teacher-assigned scores. Though scores did rise, as evidenced by the difference in mean scores for these weeks, statistically significant ($p<.05$) gains in scores were not found.

The fact that mean scores did not show significant gains over time, while disappointing, is not particularly surprising. It was unlikely that, over a ten-week class with nine writing assignments and a sample size of seven students, enough improvement could be made to detect a significant gain using a TOEFL scale. Given that mean scores between the beginning and end of the study did rise, it is possible that with a larger sample size or using a different rating scale, a significant difference could be detected. Additionally, consistently scoring a full point higher on the TOEFL writing section is quite a large gain, which likely few students could accomplish with only nine weeks of part-time study. It should also be noted that while it is possible that any gain is due to the effect of self-, peer-, and teacher-assessment, it is also possible that increased test familiarity with the TOEFL writing section, or simply increased writing practice, accounted for the change in scores. An experimental study, perhaps with a control group and conducted over a longer period of time, might provide more insight.

Inter-rater and inter-coder reliability measures all yielded statistically significant results. Inter-rater reliability measuring the consistency in the scores assigned between the teacher and the independent rater yielded a correlation coefficient of .800, which was statistically significant ($p<.05$). Inter-coder reliability, measuring the consistency in categorizing feedback according to the rubric aspect, was statistically significant ($K=.794$, $p<.05$), as was inter-coder reliability measuring the consistency in categorizing feedback as either used or unused ($K=.880$, $p<.05$). These data provide evidence that the teacher's ratings and coding according to the TOEFL rubric were consistent with that of other trained ESL teachers and a fairly reliably measure.

Correlation coefficients were also calculated between self-, peer-, and teacher-assigned scores. While there was no statistically significant correlation between self-assigned and peer-assigned scores, or between self-assigned and teacher-assigned scores, peer-assigned and teacher-assigned scores were significantly correlated, a finding consistent with Saito and Fujita's (2004) results. Though the strength of the relationship is not particularly strong at 0.437, it reached statistical significance ($p<.05$), as seen in Table 2, below. These correlation indices seem to indicate that while students could not accurately or reliably score themselves, they could more accurately and reliably score their peers. This is an interesting finding given that 67% of the survey respondents indicated that peer-evaluation was more difficult than self-evaluation.

Table 2
Score Correlation by Source

	Self	**Peer**	**Teacher**
Self	1		
Peer	0.144	1	
Teacher	0.158	.437*	1

*$p<.05$

Data relating to the use and disuse of suggestions for improvement were then analyzed by source using a chi-square test. As seen below in Table 3, feedback that was used in the revision process was compared to unused feedback across the three sources. The chi-square test performed was insignificant ($\chi^2_{(2)}=3.78$, $p>.05$), indicating that there was no significant association between use of feedback and source of feedback. A 2x2 chi-square matrix isolating

only peer and teacher feedback was also analyzed and similarly yielded insignificant results ($\chi^2_{(1)}$=3.26, corrected for continuity, p>.05), indicating that feedback given by a teacher was not significantly used over that of a peer.

Table 3
Frequency of Used and Unused Feedback by Source

	Self	Peer	Teacher	Marginal Total
Used	25	25	62	112
Unused	25	39	52	116
Marginal Total	50	64	114	228

$\chi^2_{(2)}$=3.78, relationship is insignificant (p>.05)

Insignificant chi-square tests revealing that feedback from teachers was not used significantly more than feedback from peers counters findings from Cheong (1994), Liu and Hansen (2002), and Zhang (1995) and are possibly a result of blind evaluations. Since students were not given information about who provided the feedback they received, they could not knowingly give preference to teacher feedback. This demonstrates that the perceived quality of feedback, as it related to its usability in a revision, was comparable between peers and the teacher, and could suggest that the student preference for teacher feedback found in some studies might be due to the status of the source rather than the quality of the feedback. The finding that feedback from teachers was not used significantly more than feedback from peers is consistent with Topping's (1998, 2003) assertion that there is little difference between the quality of teacher as opposed to peer feedback in a L1 context, yet may question the claims of Black et al. (2003) that students are more likely to accept feedback and criticism from peers than from teachers.

A chi-square test was also used to determine if there was a relationship between type of feedback given and its frequency of use. A chi-square test showed a significant relationship ($\chi^2_{(4)}$=13.03, p<.05), but a Cramer's V value of 0.239 revealed that while the relationship was significant, it was fairly weak. As seen in Table 4 below, it appears that whereas feedback related to use of language was used at fairly high rates, feedback related to accuracy, explanation, integration, addressing the topic, and presenting main points often went unused.

Table 4
Frequency of Used and Unused Feedback by Rubric Aspect

	Accuracy/ Explanation	Integration	Main Points/ Addressing the Topic	Organization	Use of Language	Margin Total
Used	23	6	5	45	33	112
Unused	40	10	12	36	18	116
Margin Total	63	16	17	81	51	228

*$\chi^2_{(4)}$=13.03, p<.05, Cramer's V=.239

Use of language feedback was most often used in the revision process, which was consistent with other L2 researchers (e.g., Berger, 1990; Leki, 1990; Villamil & De Guerrero,

1998) who have shown that students most often make surface-level changes. Conversely, feedback addressing further explaining or providing more detail in writing was often not used, perhaps because of the greater effort involved in adding new sentences. Similarly, the few pieces of feedback which pointed out flaws in the presentation of main points and addressing the topic, such as staying on topic and answering the given question, often required deeper structural revisions which most often went unused. Though revising the organization of an essay can also require deep structural change, and more than half of organizational feedback given was used in revisions, used organizational feedback most often related to the addition of transition words and expanding introductions and conclusions rather than deep structural changes.

The fact that no student made deep structural changes to his or her essay is not surprising. Student revisions were not re-scored, and while practicing the revision process is a fine class activity, the TOEFL tests do not measure one's ability to revise. The interest of these students in revising was therefore far lower than their interest in increasing the quality of their initial essays. It is possible, however, that if revisions had been scored again, more in-depth revisions would have been made.

Further chi-square tests were used to determine if the teacher and the students differed in the type of feedback they provided. A chi-square test run on the frequency counts in Table 5, seen below, indicated that students and teachers did significantly differ in the nature of the feedback they gave ($\chi^2_{(4)}$=10.66, p<.05), though the strength of the relationship was quite weak (Cramer's V=.216). Unfortunately, further chi-square analysis separating student feedback into feedback given through self-assessment and peer-assessment was not possible because assumptions of expected counts were not met due to low amounts of feedback given in the areas of integration, presenting main points, and addressing the topic.

Table 5
Frequency of Source Feedback by Rubric Aspect

	Accuracy/ Explanation	Integration	Main Points/ Addressing the Topic	Organization	Use of Language	Margin Total
Student	21	8	11	44	30	114
Teacher	42	8	6	37	21	114
Margin Total	63	16	17	81	51	228

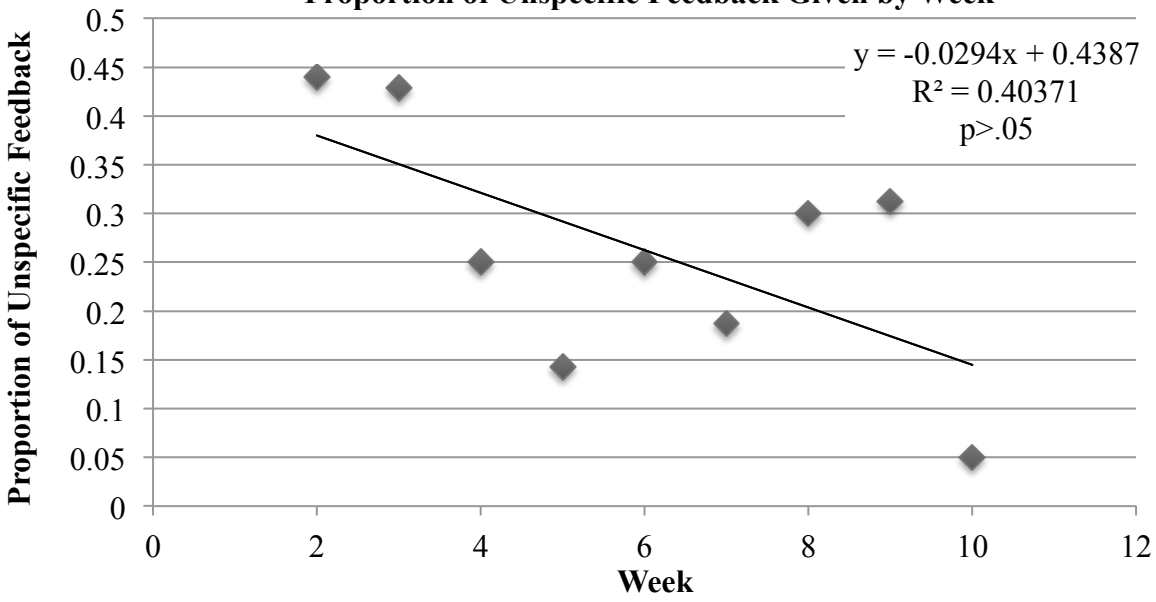
*$\chi^2_{(4)}$=10.66, p<.05, Cramer's V=.216

The largest difference between teacher and student feedback was in the area of accuracy and explanation. The teacher provided twice as much feedback as the students did which asked for more elaboration and idea development. While other studies (e.g., Beason, 1993; Leki, 1990; Villamil & De Guerreo, 1998) have addressed use of language as a primary difference between teacher- and student-evaluation, finding that student evaluators focused more on language use than teachers, the difference seen in this study was not as strong in this area. It is notable, however, that while teacher feedback for use of language generally addressed issues such as consistency in verb tense, register, and pronoun reference, student feedback for use of language tended to point out typos, morphological errors, or more vaguely comment on the need for better grammatical structure and more sophisticated vocabulary.

Linear regression analysis was used on different types of feedback over the course of the study to see if, for example, peer-given feedback related to organization increased over the study, or if self-given feedback on language use decreased during the study. None of the results were statistically significant. While Liu (2013) found that students tended to follow their teacher's lead and align their comments with those of the teacher, the students in this study did not follow this trend. Perhaps due to the blind-review nature of the study and the fact that students could not knowingly differentiate between peer and teacher feedback, students concentrated feedback in different areas than the teacher and the rates of feedback given in different areas did not significantly change throughout the study.

Linear regression analysis was also run regarding the amount of unspecific feedback that students gave both to themselves and each other throughout the study. It was hoped that the unspecific feedback students gave would decrease as the study progressed, so linear regression analysis on the proportion of unspecific feedback as a part of all student feedback given was analyzed and charted, as seen below in Figure 2. While the amount and rate of unspecific feedback given by the students did decline, regression analysis indicated that it did not decline in a statistically significant manner. Given this promising trend, however, it is possible that a larger sample size would yield significant results.

Figure 2
Proportion of Unspecific Feedback Given by Week



Similarly, proportions of feedback use according to self, peer, and teacher sources were analyzed with linear regression analysis to determine if, for example, the use of self-given feedback increased during the study, or if the use of teacher-given feedback decreased. These regression analyses showed that proportions of feedback used by source were largely inconsistent through the study, and none of these regression analyses yielded statistically significant results.

Survey questions asking students to score how much they perceived their writing to have improved in each respective rubric area were analyzed. These Likert-scale items were converted on a scale from 0 to 5 points, with 0 being no improvement, and 5 being a great deal of improvement. As seen below in Table 6, mean scores for all rubric areas fell between 3, or

somewhat improved, and 4, much improved. The highest means were in areas of addressing the topic ($\bar{x}$=3.83) and presenting main points ($\bar{x}$=3.67), while the lowest means were in the areas of accuracy ($\bar{x}$=3.00), integration ($\bar{x}$=3.33), and use of language ($\bar{x}$=3.33). Further repeated-measures ANOVA analysis revealed no significant difference ($F_{(6,30)}$= 4.706, p>.05) in the extent to which students perceived their abilities to have increased in different areas of the TOEFL writing rubric.

Table 6

Descriptive Statistics for Perceived Level of Improvement by Rubric Area

Area of Improvement	Mean	Median	St. Dev
Organization	3.50	3.50	1.05
Use of Language	3.33	3.50	1.21
Accuracy	3.00	3.00	1.10
Explanation and Elaboration	3.50	3.50	0.55
Addressing the Topic	3.83	3.50	0.98
Presenting Main Points	3.67	4.00	1.51
Integration	3.33	3.00	0.52

Survey questions asking students to score their attitudes regarding peer- and self-assessment were then analyzed. These Likert-scale items were converted on a scale from 0 to 4 points, with 0 being not at all, and 4 being a lot. As shown in Table 7 below, student attitudes were more favorable toward peer-assessment than toward self-assessment. A further t-test revealed that students liked peer-assessment significantly more than self-assessment (t(5)=2.712, p<.05) and that they perceived that they had learned significantly more from peer-assessment than from self-assessment (t(5)=2.739, p<.05). Though mean scores for the extent to which students felt peer-/self-assessment helped their revision process, and student confidence and comfort in assessing peers/themselves were higher for peer-assessment than for self-assessment, these differences did not reach statistical significance. Despite the fact that students responded more favorably to peer-assessment, 67% of students still felt that self-assessment helped their writing.

Table 7

Descriptive Statistics of Student Attitudes Regarding Peer- and Self-Assessment

Peer-Assessment				Self-Assessment			
Survey Question	Mean	Median	SD	Survey Question	Mean	Median	SD
Liked Peer-Assessment	3.50	3.50	0.55	Liked Self- Assessment	2.67	3.00	1.03
Felt peer-assessment helped revision process	3.67	4.00	0.52	Felt self-assessment helped revision process	3.00	3.50	1.26
Learned something from peer-assessment	3.67	4.00	0.52	Learned something from self-assessment	2.67	3.00	1.03
Felt comfortable assessing peers	3.50	3.50	0.55	Felt comfortable assessing themselves	3.17	3.00	0.41
Felt confident in ability to assess peers	3.00	3.00	0.89	Felt confident in ability to assess themselves	2.67	2.50	0.82

Even though all students felt that peer-assessment helped their writing, half the students commented that they questioned the correctness of peer feedback; a Russian student mentioned that "some peer grammar corrections were wrong" while another Turkish student questioned "how would I know that my peer's evaluation is correct?" These concerns echoed those felt by the students in Foley's (2013) study and have certainly been heard before from other researchers and classroom teachers; despite some negative aspects of peer-assessment, however, many positive aspects also emerged.

While students in Foley's (2013) study believed peer-assessment was more valuable for the peer assessor than for the person being assessed, students in the present study did not believe as such. Students in the present study were equally divided between those who thought peer-evaluation was more beneficial to the one who received it, and those who thought it was equally beneficial to both parties. Though no students thought that peer-evaluation was more beneficial for the one who gave the feedback, the survey responses showed that students valued both giving and receiving feedback. A German student commented that "it was nice to read another essay on the same topic and compare my ideas with the peer's," and another Turkish student remarked that peer-assessment forced her to "see the positive and the negative points, to score them, and to try to see grammar errors in somebody's essay." Some students mentioned noticing peer language, one German student indicating that she "noticed mistakes [in others' essays] that I want to prevent in the future, and good expressions that I want to use in the future," while another Turkish student said he "tried to use idioms, words, and sentences that my friends used and I liked." Those who commented on the value of receiving peer feedback said they enjoyed seeing their own mistakes from someone else's perspective and that they were able to realize the weak points in their writing and correct many of their grammatical errors. Others commented that they were able to turn peer feedback into better writing, most notably through better organization and elaboration of ideas.

Though studies have cited problems with peer-assessment with respect to leniency or low variance in scoring (e.g., MacLeod, 1999; Matsuno, 2009) and discomfort with assessing peers (e.g., Topping, Smith, Swanson, & Elliot, 2000), students in this study had no problem with peer-assessment. They reported feeling comfortable assessing their peers and, though the scores peers gave were consistently higher than the scores of the teacher, they were correlated with teacher scores, a finding consistent with Saito and Fujita's (2004) study. As Zhao (1998) also explained, blind peer review (i.e., the fact that writers did not know who their reviewers were, and reviewers did not know whose essays they were evaluating) may have been contributing factors in student comfort with the peer-assessment process.

Survey data also revealed that more than half of the students who completed the survey explicitly mentioned the value of using and becoming familiar with the TOEFL writing rubrics in both peer- and self-assessment. A Japanese student mentioned the value of checking her essays with the rubrics, and others similarly noted that a benefit was noticing different aspects of good essays by assessing them through the rubrics.

Students were also asked about the quality of their reviewers and themselves as reviewers. Although students felt their peers were better reviewers than themselves, as seen in Table 8 below, a t-test showed that this difference was not significant (t=1.168, p>.05).

Table 8
Descriptive Statistics of Student Attitudes Regarding Reviewer Quality

Survey Question	Mean	Median	SD
Felt your peers were good and reliable reviewers	3.67	4.00	0.52
Felt you were a good and reliable reviewer to your peers	3.17	3.00	0.75

Additional survey data showed that most students (83%) thought that all or most of the peer feedback they received was of good quality and that the feedback they received from peers was more helpful than the feedback they received from themselves during self-assessment. Only 17% of students thought the feedback they received during self-assessment was as helpful as the feedback they received from peers.

Some interesting findings emerged in the survey with respect to the most and least helpful types of feedback, and how students perceived their writing to have improved. While Liu and Hansen (2002) found that teachers perceive the most helpful comments in peer-assessment to be those addressing global issues such as organization, students in this study disagreed, at least ostensibly. In this study, 67% of students cited organization as the least helpful type of feedback and 83% claimed that the feedback they most frequently did not use was either explanation or organization. Interestingly, however, explanation and organization were explicitly mentioned by half of the students as specific ways in which their writing improved. Furthermore, organization received more feedback than any other category and accounted for 35.1% of all feedback given, and explanation and organization together accounted for over 63% of all substantive given feedback, as previously presented in Table 5. Mean scores for improvement in the areas of explanation and organization, however, as previously shown in Table 6, were not particularly lower or higher than other rubric areas. In contrast, while mean scores for presenting main points and addressing the topic were highest in the group, these two categories only accounted for 7.5% of all feedback given. Furthermore, there was no mention of either presenting main points or addressing the topic in student responses answering the question of how they thought their writing improved over the course of the study. Though 83% of students thought use of language was the most helpful type of feedback, and it had the highest rate of use among all rubric aspects, the mean improvement score for use of language ranked behind four other types of feedback. These are interesting findings because they may be showing some disconnect between feedback that is given, feedback that is used, feedback that is perceived to be helpful, and perceived writing improvement.

These findings could indicate that, at least with these students, even though feedback was not used and not perceived to be useful, it may have still been capable of improving writing, or at least the perception of writing improvement. This could be evidence of the "sleeper effects" that Topping (2009) indicated were possible with peer- and self-assessment, or these results might indicate that more heavily used feedback, such as use of language, may not influence the perception of increased ability in that respective rubric area. It is possible that some feedback, instead of having a more immediate effect that can be seen in a revision, has a more delayed effect and latent quality. A larger study using an analytic rubric might shed more light on this phenomenon.

CONCLUSIONS

Conclusions about the results of this study must be made extremely cautiously as the small sample size and student demographic are certainly not generalizable to all ESL or EFL students. Still, this small study can provide a glimpse into how students respond to different types of feedback.

The first research question in this study was: What is the nature of feedback source and feedback use among self-given feedback and blind peer- and teacher-given feedback? Findings revealed that students in this study did not choose to use teacher feedback significantly more than feedback from themselves or their peers. This is a likely result of the blind review process and the fact that students could not knowingly discriminate between feedback received from the teacher and feedback from a peer. Students did, however, give different types of feedback than the teacher, as the teacher gave much more explanation, elaboration, and accuracy feedback than the students. Students also used feedback differently, as they tended to favor feedback related to use of language and neglect feedback related to explanation, elaboration, accuracy, presenting main points, addressing the topic, and integration. The type of feedback students used and the amount of feedback they used by source did not change through the study.

The second research question in this study was: What is the nature of student perception of improvement as it relates to rubric areas, self-given feedback, and blind peer- and teacher-given feedback? Findings showed that although overall scores on essays did not significantly rise, students did perceive their abilities to have increased both overall and in each area of the scoring rubric. Students also responded more positively to peer-assessment than self-assessment and mentioned that both giving and receiving peer-assessment was beneficial to their writing, though they were more confident in the abilities of their peers than in their own. Surprisingly, students also perceived their abilities to have increased even in areas in which the feedback they received was not used and not regarded as useful. Students also gave the most feedback dealing with organization, the type of feedback that most students thought was the least helpful. Conversely, the highest perceived gains were in areas which accounted for the lowest amounts of feedback given.

More studies, perhaps ones using analytic rubrics, would be useful to understand the complexities of how different feedback is perceived, used, and related to writing improvement according to specific rubric areas. Additionally, further studies separating the effects of giving and receiving feedback could provide more insight into the complexities of peer-assessment.

REFERENCES

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27*(5), 427–441.

Beason, L. (1993). Feedback and revision in writing across the curriculum classes. *Research in the Teaching of English, 27*, 395-422.

Berger, V. (1990). The effects of peer and self feedback. *The CATESOL Journal, 3*, 21-35.

Birjandi, P., & Hadidi Tamjid, N. (2012). The role of self, peer and teacher assessment in promoting Iranian EFL learners' writing performance. *Assessment & Evaluation in Higher Education*, *37*(5), 513–533.

Black, P., Harrison, C., Lee, C., Marshall, B., & William, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, England: Open University Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8–21.

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London, England: GL Assessment.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (Eds.). (1971). *Handbook of formative and summative evaluation of student learning*. New York: McGraw-Hill.

Breland, H., Bridgeman, B., & Fowles, M. E. (1999). Writing assessment in admission to higher education: Review and framework (ETS Research Rep. No. 99-03). Princeton, NJ: ETS.

Chamot, A. U., & O'Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive language learning approach*. Reading, MA: Addison Wesley.

Cheong, L. K. (1994). Using annotation in a process approach to writing in a Hong Kong classroom. *TESL Reporter, 27*(2), 63-73.

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In H. L. Andrade and G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3-17). New York: Routledge.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), 328-338.

Connor, U., & Asenavage, K. (1994). Peer response groups in ESL writing classes: How much impact on revision? *Journal of Second Language Writing, 3*(3), 257-276.

Council of Europe. (1981). *Modern languages (1971–1981)*. Strasbourg, France: Council of Europe.

Council of Europe. (1988). *Evaluation and testing in the learning and teaching of languages for communication*. Strasbourg, France: Council of Europe.

Educational Testing Service. (2011). Reliability and comparability of TOEFL iBT scores. *TOEFL iBT Research, 1*(3). Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf

Elbow, P. (1973). *Writing without teachers*. New York, NY: Oxford University Press.

Falchikov, N., & Boud, D. J. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430.

Foley, S. (2013). Student views of peer assessment at the International School of Lausanne. *Journal of Research in International Education, 12*, 201-213. doi: 10.1177/1475240913509766

Hansen Edwards, J. G. (2014). Peer assessment in the classroom. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp.1-21). Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781118411360.wbcla002

Hyland, K., & Hyland, F. (2006). *Feedback in second language writing: Contexts and issues*. Cambridge, England: University Press.

Jacobs, G., & Zhang, S. (1989). Peer feedback in second language writing instruction: Boon or bane? Paper presented at the Annual Meeting of the American Educational Research Association, March 27-31. San Francisco, CA.

Leki, I. (1990). Potential problems with peer responding in ESL classes. *The CATESOL Journal, 3*, 5-19.

Lindblom-Ylänne, S., Pihlajamäki, H., & Kotkas, T. (2006). Self-, peer-, and teacher-assessment of student essays. *Active Learning in Higher Education, 7*(1), 51-62. doi: 10.1177/1469787406061148

Liu, J. (1997). A comparative study of ESL students' pre-/post conceptualizations of peer review in L2 composition. Paper presented at the 31st annual TESOL convention, March 11-15, Orlando, FL.

Liu, J. (2013). Peer response in second language writing. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-5). Oxford, England: Blackwell. doi: 10.1002/9781405198431.wbeal0902

Liu, J., & Hansen, J. (2002). *Peer response in second language writing classrooms*. Ann Arbor: University of Michigan Press.

London, M., & Tornow, W. W. (1998). 360-degree feedback: More than a tool! In W. W. Tornow, M. London, & CCL Associates (Eds.), *Maximizing the value of 360-degree feedback: a process for successful individual and organizational development* (pp. 1-8). San Francisco, CA: Jossey-Bass.

Lu, R. L., & Bol, L. (2007). A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning, 6*(2), 100−115.

Luoma, S. (2013). Self-Assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-5). Blackwell. doi: 10.1002/9781405198431.wbeal1060

MacArthur, C. A. (2007). Best practice in teaching evaluation and revision. In S. Graham, C. MacArthur, & J. Fitzgerald (Eds.), *Best practice in writing instruction* (pp. 141-162). New York: Guilford.

MacLeod, L. (1999). Computer-aided peer review of writing. *Business Communication Quarterly, 62*(3), 87-94.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75-100. doi: 10.1177/0265532208097337

Nelson, G. L., & Carson, J. G. (1998). ESL students' perceptions of effectiveness in peer response groups. *Journal of Second Language Writing, 7*(2), 113-131.

Oscarson, M. (2014). Self-Assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 712-729). Hoboken, NJ: John Wiley & Sons.

Saito, H. & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*(1), 31-54. doi: 10.1191/1362168804lr133oa

Sullivan, K., & Hall, C. (1997). Introducing students to self-assessment. *Assessment and Evaluation in Higher Education, 22*(3), 289–305.

Topping, K. J. (1998). Peer assessment between students in college and university. *Review of Educational Research, 68*(3), 249–276.

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. S. R. Segers, F. J. R. C. Dochy, & E. C. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 55–87). Dordrecht, The Netherlands: Kluwer Academic.

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology, 25*(6), 631–645. doi: 10.1080/01443410500345172

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*, 20-27. doi: 10.1080/00405840802577569

Topping, K. J., & Ehly, S. (Eds.). (1998). *Peer-assisted learning*. Mahwah, NJ: Lawrence Erlbaum Associates.
Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25*(2), 149-169. doi: 10.1080/713611428
Villamil, O. S., & De Guerrero, M. C. M. (1998). Assessing the impact of peer revision on L2 writing. *Applied Linguistics, 19*(4), 491–514.
Weaver, M. E. (1995). Using peer response in the classroom: Students' perspectives. *Research and Teaching in Developmental Education, 12,* 31–37.
Yagelski, R. P. (1995). The role of classroom context in the revision strategies of student writers. *Research in the Teaching of English, 29*(2), 216-238.
Zhang, S. (1995). Reexamining the affective advantage of peer feedback in the ESL writing class. *Journal of Second Language Writing, 4*(3), 209-222.
Zhao, Y. (1998). The effects of anonymity on computer-mediated peer review. *International Journal of Educational Telecommunication, 4*(4), 311−345.

APPENDIX A

Self-Evaluation Form
Independent Tasks

After you complete your essay, please read it and complete the following questions:

TOEFL scores for the Independent Writing task are based on the following aspects of writing:
Addressing the topic—(the extent to which the question was answered fully and completely)
Organization—(the extent to which the ideas logically progress and connect to each other)
Explanation and elaboration–(the extent to which there were enough details to fully explain the position)
Use of language—(the extent to which grammar or vocabulary were used correctly)

1) In your opinion, which of these aspects is the strongest in your essay? (please choose only one)

2) In your opinion, which of these aspects is the weakest in your essay? (please choose only one)

3) Using the rubric for Independent Tasks, which score (from 0-5) would you give yourself?

4) Please provide one or two specific examples of how your essay can be improved.
NOTE: This does NOT mean "what are some general things that can be done to become a better writer?" (such as *read more* or *increase my vocabulary*) it more closely means "what's wrong with this specific essay, and what can I do about it?" (such as *give more examples in the second paragraph* or *make the introduction clearer*)
1)
2)

APPENDIX B

Peer-Evaluation Form
Integrated Tasks

After you read the essay, please complete the following questions:

Presenting main points—(the extent to which important and relevant information is selected)
Accuracy—(the extent to which the main points are presented coherently and accurately)
Integration—(the extent to which the information in the reading is presented in relation to the information in the lecture)
Organization—(the extent to which the ideas logically progress and connect to each other)
Use of language—(the extent to which grammar or vocabulary were used correctly)

In your opinion, which of these aspects is the strongest in this person's essay? (please choose only 1)

In your opinion, which of these aspects is the weakest in this person's essay? (please choose only 1)

Using the rubric for Integrated Tasks, which score (from 0-5) would you give this person?

Did you notice any grammar mistakes in this person's essay? If so, what kind of mistakes?

Please provide one or two specific examples of how this person's essay can be improved. NOTE: This does NOT mean "what are some general things that can be done to become a better writer?" (such as *read more* or *increase my vocabulary*) it more closely means "what's wrong with this specific essay, and what can he/she do about it?" (such as *give more examples in the second paragraph* or *make the introduction clearer*)

1)
2)

APPENDIX C

Revision Evaluation Form

First, cut and paste your original essay here:

Now, go back and revise things in your essay. Please either use the Track Changes feature (under the Review tab in MS Word), or use a different color for things that you change, like this:

I think~~s~~ parents are the ~~better~~ best teachers. I think this because parents are the first people who know you and because they raise you. They are the first people who teach you right from wrong and who love you for many years before you ever go to school. In this essay I will explain why parents are the best teachers.

Second, as you change things in your essay, cut (or copy) and paste the feedback that you used into the appropriate boxes. Things that you change in your essay that help it go under #1.

Lastly, when you finish, look at the feedback you didn't use and decide why you didn't use that feedback. If you thought it was good, but you just didn't use it, put it under #2. If you thought it wasn't good feedback and wouldn't help your essay, put it under #4. Please include feedback from your self-evaluation as well as feedback from others in the form below.

1) Feedback that I thought was good, and I incorporated it into my revision:
(Liked and used)

2) Feedback that I thought was good, but I didn't incorporate it into my revision:
(Liked but didn't use)

3) Feedback that I thought was not very good, but I still incorporated it into my revision:
(Didn't like but used)

4) Feedback that I thought was not very good, and I didn't incorporate it into my revision
(Didn't like and didn't use)

Other comments:

APPENDIX D

Student Survey

At the end of your first essay each week, you were asked to complete a self-evaluation. The self-evaluation asked you to indicate a positive point, a negative point, a score, and ways to improve. This section will ask you about your experience with self-assessment.

1. What were the positive aspects of doing self-assessment?
2. What were the negative aspects of doing self-assessment?
3. Do you feel that self self-assessment helped your writing?
4. If you answered yes to the previous question, How or in what ways do you feel your writing improved because of self-assessment?

After you submitted your essay, you received a peer's essay and were asked to review it. The peer evaluation asked you to indicate a positive point, a negative point, a score, grammar errors, and ways to improve. This section will ask you about your experience with giving peer-assessment.

5. What were the positive aspects of giving peer-assessment?
6. What were the negative aspects of giving peer-assessment?
7. Do you feel that peer-assessment helped your writing?

8. If you answered yes to the previous question, How or in what ways do you feel your writing improved because of giving peer-assessment?

After you reviewed a peer's essay, you received feedback from a peer, the teacher, and your self-evaluation. In this section, please answer questions about the feedback you received.

9. Describe the quality/helpfulness of the feedback you received.
10. Was the feedback you received from peers generally more/equally/ or less helpful compared to the feedback you gave yourself in your self-evaluations?
    [If you thought peers gave you better feedback than your self-evaluation, choose more helpful; if you thought they were about the same, choose equally helpful; if you thought your self-assessments were better than peer feedback, choose less helpful]
11. What type of feedback did you find most useful in your revision process?
    example: language, elaboration, organization, integration, main points
12. What type of feedback did you find least useful?
    example: language, elaboration, organization, integration, main points

After you received feedback, you were asked to revise your original essay. This section will ask about your experience revising your essay.

13. What type of feedback did you usually use in your revision?
    example: language, elaboration, organization, integration, main points
14. What type of feedback did you usually NOT use in your revision?
    example: language, elaboration, organization, integration, main points

These next questions will ask you about your overall experience as a participant in the study.

15. Overall, do you think your writing improved over the course of the study?
16. If you answered yes to the previous question, how or in what ways did your writing improve?
17. Please indicate how your writing improved in the area of organization *(in both independent and integrated tasks)*
18. Please indicate how your writing improved in the area of use of language *(in both independent and integrated tasks)*
19. Please indicate how your writing improved in the area of addressing the topic *(in independent tasks)*
20. Please indicate how your writing improved in the area of explanation and elaboration *(in independent tasks)*
21. Please indicate how your writing improved in the area of presenting the main points *(in integrated tasks)*
22. Please indicate how your writing improved in the area of accuracy *(in integrated tasks)*
23. Please indicate how your writing improved in the area of integration *(in integrated tasks)*

*These next few questions will address **self**-assessment*

24. In general, how comfortable did you feel reviewing your own essays through self-evaluations?
25. In general, how confident did you feel in your ability to evaluate your own work?
26. Do you feel that self-evaluation helped your revision process?
27. Do you like self-assessment?

28. Do you feel like you learned something from self-assessment?

*These next few questions will deal with **peer**-assessment.*

29. In general, how comfortable did you feel reviewing peers' essays through peer-evaluation?
30. In general, how confident did you feel in your ability to evaluate the work of your peers?
31. Do you feel that peer-evaluation helped your revision process?
32. Do you like peer-assessment?
33. Do you feel like you learned something from peer-assessment?
34. Do you feel that your peers were good and reliable reviewers?
35. Do you feel that you were a good and reliable reviewer for your peers?
36. Who do you think peer-evaluation is more beneficial for? The person giving the feedback or the person receiving it?
37. Which was more difficult for you, peer-assessment or self-assessment?
38. Did you learn more from your own self-evaluation feedback or from peers' feedback?